THE CHOICE BETWEEN TWO HYPOTHESIS TESTS

Erik Ruist
(Stockholm School of Economics, Stockholm, Sweden)

SSE/EFI Working Paper Series in Economics and Finance
No.667
June 2007

Abstract

Most standard hypothesis tests have high power only against a limited space of alternative hypotheses. With the advent of new tests for the same hypothesis, claimed to have higher power for some alternatives, but lower power for other, the practitioner often has to make a choice between two alternative tests. This paper recommends the use of a pre-test to guide this choice, or the combined use of both tests.

Key words: sign test; t test; two-dimensional criterion

JEL clasification: C12

1. Introduction

Many classical statistical hypothesis tests are based on rather rigid assumptions regarding the distribution under scrutiny, for example that it is normal. Since empirical evidence often shows that this property can be doubted, new tests have been developed which are optimal when the distribution is non-normal. The existence of several alternative tests for one and the same hypothesis naturally leads to a problem of choice between them in a practical situation.

As part of a description of a suggested new test, its power is usually calculated for the space of alternative hypotheses. This space is sometimes multi-parametric, which makes it difficult or impossible to derive the power as an analytic function of all parameters, so it is often evaluated by simulation for a limited number of points. This is the case we are going to investigate.

Ordinarily, it turns out that the new test has higher power than the traditional one for some combinations of parameter values, while the opposite is true for other combinations. Thus, the decision on which test to use in a practical situation is not easily made. Existing criteria for comparing two tests or for finding an optimal one usually require the knowledge of the power function of each test over the whole hypothesis space, and are consequently not directly applicable in the present case. In the following, we will suggest three possible ways to act in a situation of this kind. First, we will modify some existing criteria to make them applicable also in the present case. As these modifications may be critizised for having a large subjective element, we have two alternative solutions.

A natural thought in this situation is to take advantage of both tests, preferably by using the most powerful one in each situation. There may exist a test which indicates whether the present distribution belongs or not to a subclass of alternatives where one of the two tests has low power. This preliminary test can then be used as a mechanism for choosing between the alternative tests, and the expected power of the combined test is probably for each distribution close to that of the better one of the two alternative tests. Such a procedure is used as one alternative in testing Granger noncausality by Péguin-Feisolle and Teräsvirta (1999).

In cases where no such preliminary test is available, one may choose between the alternative tests by a pure random mechanism with a pre-set probability. The expected power would then for all cases lie between those of the alternative tests, and may be better than both according to some criteria. It may, however, not seem to be very rational to let a random process decide the choice of test. Instead, as a third solution, we suggest the use of *both* tests. Thus, the decision would be to reject the null hypothesis if at least one of the alternative tests rejects it. Of course, the level of such a combined test has to be modified compared to the original tests. If these have the size 5 per cent, the combined test will have a size between 5 and 10 per cent, if the critical values of the test statistics are unchanged. Thus, these have to be modified in order to give the combined test the desired size. This can of course be done by changing the critical value of one or the other, or both, tests. How this should be done in an optimal way is a problem to be solved. The power of such a combined test will in most cases, but not necessarily, fall between those of its components.

2. Choice of an optimal test

The problem of choosing between several tests, or rather finding a test that is optimal in some sense has been discussed at least since the 1930s. Let us introduce the following notations:

The distribution or Data Generation Process F under consideration is known to belong to a class Ω. We want to decide to which subclass ω_i of Ω that F belongs. Normally, only two subclasses are considered, so that $\omega_1 \cup \omega_2 = \Omega$. In many cases, ω_1 consists of only one element – a *simple* hypothesis – but we shall not make this restriction. Let us denote the decision to accept $F \in \omega_i$ by d_i, i=1,2. To make this decision, we use a test φ, which for every sample point x indicates the probability with which decision d_2 should be made. To achieve this, the test φ is associated with a function $\psi(x)$, the test criterion, by the relation

$$\varphi(x) = \begin{cases} 1 \text{ if } \psi(x) > c \\ q \text{ if } \psi(x) = c \\ 0 \text{ if } \psi(x) < c \end{cases} \qquad (1)$$

where c and q are parameters, $0 \le q \le 1$. Often, $P(\psi(x) = c) = 0$, and the value of q is irrelevant.

The decision we take on the basis of $\varphi(x)$ may be wrong, but we want of course to minimize the probability of a wrong decision. In accordance with common practice, we denote

$$\alpha = \sup_{\omega_1} P(d_2 \mid \omega_1) = \textit{size of the test}$$

$$P(d_1 | \omega_2) = 1 - \textit{power}$$

Thus, while the test´s size is unique, its power is a function of the distribution F.
Now, in order to find an optimal test or to compare two tests, we can either look at their performance over the whole of Ω, or subjectively choose a value of α and investigate the power over various points of ω_2. We will mainly discuss this second procedure.

As was noted already by Neyman and Pearson (1936), there may exist a Uniformly Most Powerful test, i.e. one test that has higher power than all other tests over the whole range of ω_2. This is, however, true mainly for rather restricted Ωs, e.g. normal distributions. In the more common case, when one test has higher power in one part of ω_2, and a different test has higher power in the other part, additional criteria have to be used.

Wald (1942) suggested that one should evaluate for each $F \in \omega_2$ the highest power that is achieved by any test under consideration, and then choose the *most stringent* one, i.e. the test φ' with the smallest maximal deviation from this envelope power function. This test will thus minimize

$$\max_{F \in \omega_2} \left[\sup_{\varphi} P(d_2 | \varphi, F) - P(d_2 | \varphi', F) \right]$$

To our knowledge, this criterion has seldom, if ever, been used in practical applications. We will, however, return to it in a later section.

Most other criteria for selection of an optimal test require that a weight $W(F)$ is attached to each F in ω_2, and sometimes also in ω_1. These weights may be interpreted either as prior probabilities within Ω, or as indications of the seriousness of, or loss incurred by, a wrong decision, given that the true distribution is F. Lindley (1953) included both probability and loss in his interpretation of the weights, and was thus able to compute an expected loss for every test. Then of course the test with the lowest expected loss is the optimal one. Due to the difficulty in determining such weights, also this criterion is seldom used.

In connection with his work on statistical decision functions, Wald (1950) also weighted the Fs, but interpreted the weights only as indications of the losses from a wrong decision. Thus, for any $F \in \omega_2$, the expected loss is

$$r(\varphi, F) = P(d_1 | \varphi, F) \cdot W(F)$$

In fact, Wald also included the expected loss from an error of the first kind, i.e. the error of choosing d_2 when $F \in \omega_1$, but we may, following Hoeffding (1951), disregard this complication here.

For any test φ, the distribution $F' \in \omega_2$ that maximizes the expected loss, so that

$$r(\varphi, F') = \max_{\omega_2} r(\varphi, F)$$

may be called the *least favorable distribution* with respect to φ. The test φ' is said to be of *minimax risk* if its maximal risk is smaller than that for all other tests, i.e.

$$r(\varphi', F) = \min_{\varphi} \max_{\omega_2} r(\varphi, F)$$

The minimax risk test is thus optimal for its least favorable distribution. This criterion for the choice of a test will be further discussed below.

In order to apply any of the criteria discussed above, it is clear that we need an evaluation of the power of each test over the whole of ω_2. For most criteria, we also need a weight function, defined over ω_2. In the situation that we describe here, there are practical obstacles to the derivation of both of these data. First, there is usually no rule to guide the allocation of weights to the various F in ω_2. Second, we calculate the power of the test only for a limited number of Fs. We shall return to this second objection in the next section.

3. Modified Choice Criteria

When investigating a new test it is important to evaluate its power throughout the whole of ω_2. In many cases, the power can be expressed as an analytic function of the parameters that characterize Ω, and thus ω_2. However, when Ω is a multi-parameter class of distributions or DGPs, this may be difficult or even impossible. Then, the praxis is to use Monte Carlo simulations to find the power for some pre-determined parameter value combinations, see e.g. Eklund (2003), Gonzalez (2004), Sandberg (2005), and Strikholm (2004). Let us denote the set of investigated points in ω_2 by ω_{20}.

In order to use one of the criteria listed above for the choice between two or more tests, it is essential to express the power as a function of the elements of ω_2. If this function were linear, or at least polynomial, it would be possible to estimate it with a regression equation based on the observed points ω_{20}. Whether or not this gives a reasonable description of the data could

be investigated with an analysis of variance, which could indicate if there are any interaction effects. If such effects are present, it seems difficult to construct a reasonable analytic expression of the power function. Apparently, other criteria for the choice between tests have to be used.

The minimax criterion, as it was originally developed by Wald, applied weights to all elements F' of Ω (ω_1 as well as ω_2), indicating the "loss" incurred by a wrong decision, given that $F=F'$. Hoeffding(1951) modified this, and looked only into ω_2, thus preferring the test (or rather the test family, with α to be determined exogenously) that has smallest maximal expected loss. It seems that a similar criterion can be used, even if the power is evaluated only for the points in ω_{20}, say F_i , i =1,…,k. It is usually not feasible to attach weights to F_i , so we have to base our comparisons on the estimated power values only. Let the power estimate of test φ_j for $F = F_i$ be $p_{ij} = P(d_2|\varphi_j, F_i)$. Then the test φ' which satisfies

$$p'_{ij} = \max_j \min_i p_{ij}$$

is the *maximin power test* in terms of the present information. Thus, if we look at the worst performance for each test within ω_{20}, the maximin power test is not so bad as the other ones. We may call the distribution F_L for which the minimum power for a test is obtained, for its *least favorable investigated distribution.*

In a similar way, we may define a test φ' to be *most stringent* for the points in ω_{20} if it has the smallest maximum deviation from the highest power observed in each point, i.e. it minimizes

$$\max_{F\in\omega_{20}}\left[\sup_{\varphi} P(d_2|\varphi, F) - P(d_2|\varphi', F)\right]$$

In spite of the fact that the stringency criterion has been little used by statisticians since it was introduced, we are inclined to prefer it to the minimax criterion. It favours a test that is a little worse than the other one for some $F \in \omega_{20}$, but much better for others. This seems to be more reasonable than to let the F with the lowest test power be decisive for the choice. We shall scrutinize the outcome of our simulations according to both criteria.

If it seems inappropriate to use the test which has the best performance in the worst possible case, whether in absolute power or in deviation from the envelope, this general idea still points to the importance of choosing Ω properly and not unnecessarily wide. It may be profitable to ask: Can we reduce Ω so as to exclude the least favorable distribution of one of the tests?

Instead of choosing between the two tests according to some criterion of the type discussed above, perhaps it would be possible to find some random mechanism for the choice. Already a 50/50 choice would produce a combined test which is in many cases better in the minimax sense and more stringent than each of its two components, since the expected power lies between those of the two original tests.

Sometimes it is possible to use a preliminary test instead of a purely random mechanism. This test may be able to indicate if we are in a part of Ω where one of the main tests generally has higher power than the other one. If this is possible, the expected power of the combined test will be higher than with a purely random choice. The results given by Péguin-Feissolle and

Teräsvirta (1999), who used this idea in testing Granger causality, show that the combined test was rather successful in terms of stringency: it was never best, but on the other hand never very far from the best among the five investigated tests.

There is a third possibility in the choice between two tests, and that is to use both of them. The test criterion would then be: reject the null hypothesis if it is rejected by at least one of the two tests. This is equivalent to using a two-dimensional rejection region. Thus, if the test criterion of the first test is ψ_1 with a rejection region $R_1 = \{\psi_1 > c_1\}$, and that of the second test is ψ_2 with rejection region $R_2 = \{\psi_2 > c_2\}$, we now use the rejection region $R = R_1 \cup R_2$. It is clear that the size of this combined test will be higher than those of the original tests. If these are both 5 per cent, the combined test will have a size of between 5 and 10 per cent, depending upon how correlated the two test criteria are. In order to return to the intended size, 5 per cent, we have to modify the critical limits of the original tests upwards, to $(c_1 + c_{11})$ and $(c_2 + c_{21})$. Several choices of c_{11} and c_{21} would yield the desired result, and the problem of finding an optimal (c_{11}, c_{21}) remains to be solved.

4. A practical illustration

As an illustration of the procedures discussed above, we have elaborated an example, which is not primarily intended to provide new information about the tests, but only to show the practical handling of the data in a specific situation.

Suppose we know that the distribution we want to investigate is symmetric about its mean, but may otherwise be of any form. We want to test the hypothesis that the mean μ is = 0 against the alternative $\mu > 0$, and we take a sample of n independent observations.

From classical theory we know that the *t*-test is uniformly most powerful, if Ω only contains normal distributions. For other Ωs , other tests may be more powerful. We shall here investigate the performance of the sign test. This test can of course only test the situation of the median of the distribution, but since we decided that Ω only contains symmetrical distributions, the median is equal to the mean.

The *t*-test and the sign test have been compared many times before. An early example is Gibbons (1964), who calculated the power of the two tests for distributions with various values of skewness and kurtosis. The comparison did, however, not result in any recommendation on which test to use, which is the ultimate goal of the present investigation.

To find out if and when the sign test outperforms the *t*-test, we have calculated the power of the two tests for a number of cases in ω_2 by Monte Carlo simulation. The choice of points in ω_2 is by no means self-evident. It is clear that the power depends on the sample size *n,* on our choice of test size α, and further on the characteristics of ω_2, i.a. the mean μ. We have estimated the power of the two tests for the following values of the parameters:

Sample size n = 25, 100
Mean μ = 0.25, 0.5
Size of the test $0.01 < \alpha < 0.20$

and for the following five different forms of distributions:

1. Rectangular distribution

2. Normal distribution
3. Logistic distribution
4. The χ_1^2 distribution, mirrored around x=0 to make it symmetrical. For simplicity, we will call this symmetrical distribution the *normal-square* distribution
5. A χ_2^2 - like distribution, similarly mirrored. Its absolute value was obtained as minus the logarithm of a rectangularly distributed variable, and we thus call it the *log-rectangular* distribution.

All distributions have been normalized to have $\sigma = 1$. To give an impression of the characters of these distributions, a frequency histogram of a sample of 1000 observations of each non-normal distribution, together with the corresponding normal curve, is given in the Appendix.

To estimate the power of the tests, 10 000 simulations have been made at each point. This gives a 5 per cent confidence interval for the power estimations of about ± 0.010 around the tabulated values for power around 0.5, and ± 0.006 for power around 0.9.

We start by establishing the critical values of the test quantities for the two tests. To begin with, we restrict ourselves to consider tests at the 5 per cent significance level. The determination of critical values is not trivial, since ω_1 contains not only one element, but the whole set of symmetrical distributions with $\mu = 0$. The size of the test is then the highest value of the probability of an error of the first kind for any element in this set. In practice, we have to restrict ourselves to the values for those distributions that we intend to investigate.

For the sign test, this does not cause any trouble. The test criterion $\psi(x)$ in (1) is here G, the ratio of positive observations. Since for all elements of ω_1 the probability of any observation to be positive is ½, the distribution of G is the same over the whole of ω_1 . We can easily find out that the probability of 17 or more positive observations in a sample of 25 is 0.054. Thus, if we reject the hypothesis $\mu = 0$ when the proportion of positive observations G is $\geq 17/25 = 0.68$, we have a test of size 0.054. In order to construct a test of size 0.050, we will have to let q in equation (1) have a value between 0 and 1 for $G = 0.68$. Even if this is not a solution that is used in practice, we use this construction to get a sign test of the same size as the t test.We thus let q be 0.880.

For the t test, the critical value is not so easily obtained. From a table of the t distribution we can state that for a normal distribution and for $n = 25$, the statistic $t = \dfrac{\bar{x}}{s/\sqrt{25}}$ has for 24 d.fr. the 0.05 critical value 1.711 . However, for other ω_1 distributions, this critical value may give higher rejection probabilities, and the size of the test is then higher than 0.05.

By simulation, we have estimated the rejection probabilities for various critical values of t for the distributions mentioned above. As a matter of fact, we have for each distribution estimated the critical limit that gives the rejection probability = 0.05. The result was the following:

Rectangular	1.732
Logistic	1.726
Log-rectangular	1.734
Normal-square	1.688

Table 1. Power of the t-test and the sign test at selected points in ω_2. n=25, α = 0.05

	μ=0.25			μ=0.5		
	t test	*sign test*	*difference*	*t test*	*sign test*	*difference*
Rectangular	*0.31*	*0.18*	*0.13*	*0.78*	*0.43*	*0.35*
Normal	*0.31*	*0.25*	*0.06*	*0.78*	*0.63*	*0.15*
Logistic	*0.33*	*0.30*	*0.03*	*0.78*	*0.70*	*0.08*
Log-rectangular	*0.35*	*0.44*	*-0.09*	*0.77*	*0.84*	*-0.07*
Normal-square	*0.39*	*0.82*	*-0.43*	*0.80*	*0.98*	*-0.18*

Now, if we choose the largest of these values, i.e.1.734, as the critical limit of the *t* test, we will for all of the investigated distributions get a rejection probability of at most 0.05, which is thus the size of the test within the investigated part of ω_1. Using these critical values, the power of the two tests was found to assume the values given in Table 1.

It is seen that the *t*-test has higher power than the sign test for the rectangular, the normal, and the logistic distribution. For the more extreme distributions the sign test is, however, more powerful. Since for $\mu = 0.25$ as well as for $\mu = 0.5$ the lowest power obtained for the *t*-test is higher than that of the sign test, the *t*-test is better than the sign test in the minimax sense for the investigated distributions.

Looking for the most stringent test, we find on the other hand that for $\mu = 0.25$, the power of the *t*- test is 0.43 lower than that of the sign test for the Normal-square distribution. The sign test is never more than 0.35 below the *t*-test. Thus, the sign test is the more stringent one.

For a further analysis of the power it would be advantageous to express it as a function of one or more parameters that characterize the distributions in ω_2. One obvious candidate parameter is the kurtosis of the distribution. To illustrate its possible influence, Figure 1 shows the power for the five investigated distributions, where these have been characterized by their obtained average kurtosis. The power of the *t*-test is relatively unaffected by the

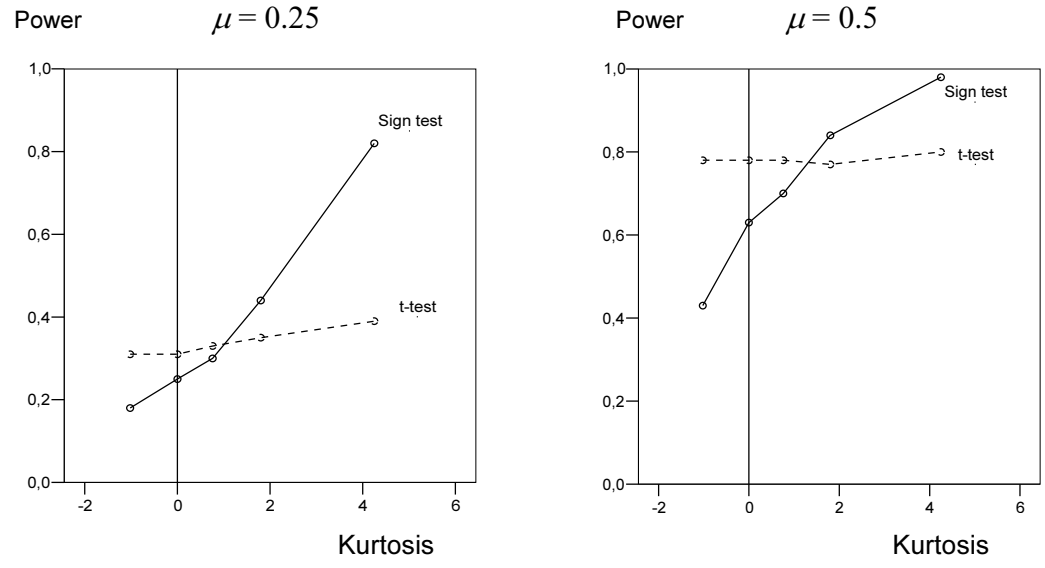


Figure 1. Power for the t-test and the sign test for distributions with different kurtosis. n = 25; α=0.05

kurtosis, while the sign test has a much higher power when the distribution has a high kurtosis. This is true for each of the two values of μ that we have investigated, even if the level of the power is different. It seems reasonable to suppose that we would find the same for other values of μ.

We should, however, also investigate whether this behaviour is independent of the sample size and of the test size. As to the sample size it should be clear that an increase by the factor four gives the same effect on the *t*-test as a decrease in μ to the half, except for the slight differences between the *t* distributions for various degrees of freedom. The same is probably approximately true for the sign test. This is confirmed by our simulations. The difference in power between the combinations ($n = 25$; $\mu = 0.5$) and ($n = 100$; $\mu = 0.25$) is at most 0.02 for the *t*-test and 0.05 for the sign test.

The influence of the test size can be shown as in Figure 2, where the power is calculated as a function of the test size for a couple of specific points in ω_2, i.e. for ($n = 25$; $\mu = 0.5$) and for the normal and the normal-square distributions. The results for the other investigated distributions (not shown here) confirm the conclusions that can be drawn from this picture. As we have noted for $\alpha = 0.05$, the power of the *t*-test is rather unaffected by the form of the distribution, while the sign test is rather sensitive. The sign test is clearly less powerful than the *t* test for the normal distribution, but more powerful for the normal-square. Since the maximum advantage of the sign test in the latter case is larger than its disadvantage in the fomer, the sign test is more stringent than the *t*-test for these cases, while the *t*-test is better according to the minimax criterion.

After these investigations we can be fairly confident that the difference between the *t*-test and the sign test that we have found are valid – although with varying magnitude – for most

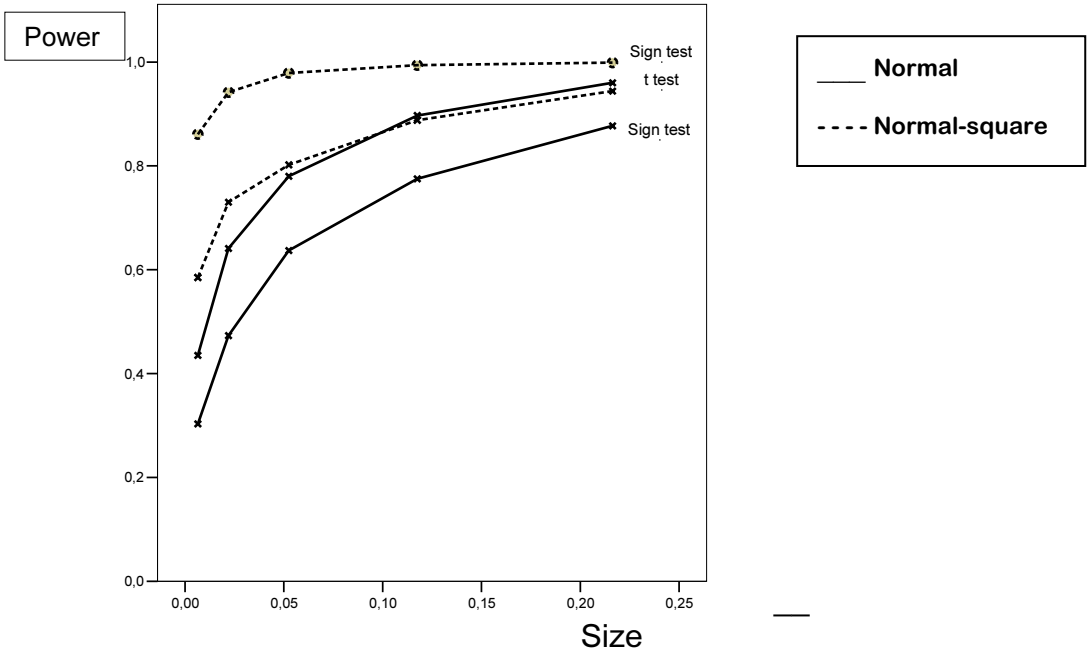


Figure 2. Power as a function of test size for n=25, μ=0.5

Table 2. Power of four tests at selected points in ω_2. n=25, a=0.05

	μ=0.25				μ=0.5			
	t test	*sign test*	*comb test*	*two-dim test*	*t test*	*sign test*	*comb test*	*two-dim test*
Rectangular	*0.31*	*0.18*	*0.31*	*0.27*	*0.78*	*0.43*	*0.78*	*0.74*
Normal	*0.31*	*0.25*	*0.31*	*0.30*	*0.78*	*0.63*	*0.78*	*0.76*
Logistic	*0.33*	*0.30*	*0.33*	*0.33*	*0.78*	*0.70*	*0.77*	*0.78*
Log-rectangular	*0.35*	*0.44*	*0.37*	*0.40*	*0.77*	*0.84*	*0.80*	*0.83*
Normal-square	*0.39*	*0.82*	*0.67*	*0.72*	*0.80*	*0.98*	*0.93*	*0.96*

values of n, α, and μ. Thus, if we could use the sign test for distributions with high kurtosis, and the *t* test in more "normal" situations, we might get a better result on average.

5. Combined tests

In order to exploit this possibility we have constructed a *combined test* in the following way: The kurtosis of the sample is calculated, and if it is less than 2, the *t*-test is used for testing whether the average is 0 or positive. If we get a kurtosis above 2, the sign test is used instead. It turns out that the size of this combined test is in fact slightly higher than 0.05. We have corrected this by setting q to 0.13 instead of 0.88 in the sign test.

It is clear that when the distribution is in fact normal or rectangular, we seldom get an observed kurtosis above 2, and the *t*-test is used in the majority of cases. Thus, the average power of the combined test is very close to that of the *t*-test for those distributions. On the other hand, for more extreme distributions, the power is close to that of the sign test, see Table 2. Note that the power of the combined test does not necessarily fall between those of its components, since the election probability is correlated with the power. In this case we have, however, not detected any such result.

According to our results for $\mu = 0.5$, the lowest power found for the combined test as well as for the *t*-test is 0.77, and for the sign test 0.43. Thus, for this μ the combined test is, together with the *t*-test, better in the minimax sense than the sign test, when the comparison is restricted to the investigated distributions. The same is true for $\mu = 0.25$.

Comparing the combined test with the best test in each investigated situation, we find that it is never more than 0.15 below (for $\mu = 0.25$, normal-square), while the *t*-test in the same case is 0.43 behind, and the sign test has 0.35 lower power than the best for $\mu = 0.5$, rectangular distribution. Thus, the combined test is without competition the best one according to the stringency criterion. It is never best, but it is seldom very far from the best!

6. A two-dimensional criterion test

The alternative solution to the choice between two tests that we suggested earlier was to use a two-dimensional rejection criterion. We constructed such a test (for $n = 25$) by increasing the critical limit for the sign test to 0.70, and then finding the *t* value that gives the combined size 0.05. As before, the limit is different for the investigated distributions, and we chose the

Table 3. Power of two-dimensional tests with different critical limits. n=25, α=0.05

	μ =0.25		μ =0.5	
	$c_G = 0.70$	$c_G = 0.74$	$c_G = 0.70$	$c_G = 0.74$
	$c_t = 1.855$	$c_t = 1.755$	$c_t = 1.855$	$c_t = 1.755$
Rectangular	*0.27*	*0.31*	*0.74*	*0.77*
Normal	*0.30*	*0.31*	*0.76*	*0.78*
Logistic	*0.33*	*0.33*	*0.78*	*0.78*
Log-rectangular	*0.40*	*0.37*	*0.83*	*0.80*
Normal-square	*0.72*	*0.60*	*0.96*	*0.92*

highest value, which was 1.855. We thus rejected the null hypothesis if $G > 0.70$ and/or $t > 1.855$. This choice is certainly arbitrary. We will discuss the choice below.

The power of this two-dimensional criterion test is rather close to that of the combined test, see Table 2. In general, the differences are not statistically significant. Both are better than the sign test according to both criteria used here, and more stringent than the *t*-test, but not better according to the minimax criterion. The two-dimensional test also beats the combined test according to the stringency, but not according to the minimax criterion. It has also the advantage that it does not require a preliminary test in order to discriminate between points in ω_2 . It can easily be constructed when we have two alternative tests for our main hypothesis. The remaining question is to find the optimal pair of values (c_{11}, c_{21}) for the changes in the critical values of the test criteria. In order to illustrate the effect of varying the limits, we have calculated the power also for a test with c = (0.74, 1.755) and compared it with the one discussed above with c = (0.70, 1.855). The result is shown in Table 3. As could be expected, the test with the larger c_G has a power closer to that of the *t*-test than that with the smaller c_G, as fewer observations are rejected because of a high G, and more because of a high t. It seems that, from a stringency point of view, the rather small improvement for the low-kurtosis distributions does not compensate for the losses for the high-kurtosis distributions.

We have certainly not solved the problem of an optimal choice of (c_{11}, c_{21}). We recommend some experimentation in order to find values that make the two-dimensional test superior to its two components according to the preferred criterion.

7. Conclusions

We have investigated the problem of choosing between two tests in a situation when the power of the tests could be computed (mainly by simulation) for only a limited number of elements ω_{20} in ω_2 . We expressed a preference for using Wald´s stringency criterion for the choice. We also suggested that both tests should be used together, either by selecting one of them by a preliminary test, or by using a two-dimensional test criterion. For the example that we used, a test of the mean in a symmetrical distribution, it turned out that the combined test and the two-dimensional criterion test performed equally well, and were more stringent than the *t*-test as well as the sign test.

If this similarity in performance of the two ways to use both tests is common for a great many other situations, it seems clear that the two-dimensional criterion test is preferable, since it

does not require the existence of a preliminary test. We thus recommend that this solution is investigated in other cases of choice between tests.

References

EKLUND, B (2003), *Four contributions to statistical inference in econometrics*. Stockholm

GIBBONS, J.D. (1964), Effect of non-normality on the power function of the sign test. *Journal of the American Statistical Association*, Vol 59, 142-148.

GONZALEZ, A (2004), *Nonlinear dynamics and smooth transition models*. Stockholm

HOEFFDING, W. (1951), "Optimum" nonparametric tests. *2nd Berkeley Symposium*, 83-92.

LINDLEY, D W (1947), Statistical inference. *JRSS B,* Vol. 15,. 30-76.

NEYMAN, J. and E.S. PEARSON (1936), Contributions to the theory of testing hypotheses, Part I, *Statistical Research Memoirs*, Vol I, London.

PEGUIN-FEISOLLE, A, and T. TERÄSVIRTA (1999), A general framework for testing the Granger noncausality hypothesis, *Stockholm School of Economics, Series in Economic and Finance,* No 343.

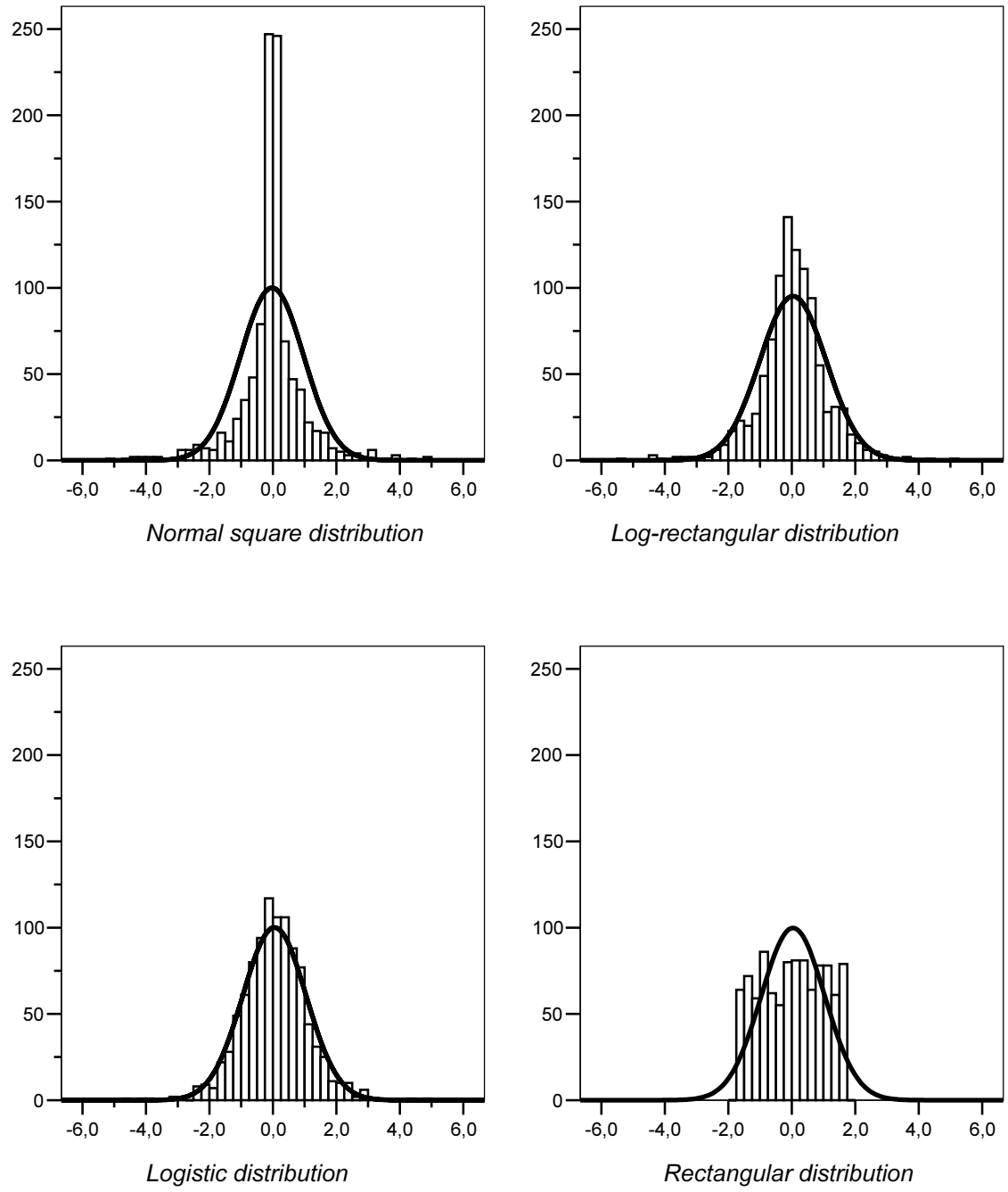
SANDBERG, R (2004), *Testing the unit root hypothesis in nonlinear time series and panel models*. Stockholm

STRIKHOLM, B (2004), *Essays on nonlinear time series modelling and hypotheis testing.* Stockholm.

WALD, A. (1942), On the principles of statistical inference. *Notre Dame Mathematical Lectures,* No 1.

WALD, A. (1950), *Statistical decision functions*. New York

Appendix



Frequency histograms of the investigated distributions. The normal distribution is inserted in all diagrams for comparison.